Server Memory Road Map

Ricki Dee Williams, Theresa Sze, Dawei Huang, Sreemala Pannala, Clement Fang

System Electrical Technology, Packaging and PCB Technology, Mixed Signal Technology, Memory Technology Group
Oracle Corporation

Server Memory Forum Shenzhen 2012
Outline

- Off-Package IO Bandwidth Historic Trend
- Oracle T3 CPU & IO
- 20Tbps IO Bandwidth Challenges
  - The Power, Area, and Frequency Wall
  - The Paradigm Shift – A Revolution
  - Potential 2D Solutions
- The Memory Wall
  - Potential 3D Solutions
- Conclusion
Total Bandwidth - Historical Trend

Bandwidth trend follows Moore Law
Double every 18 months in the past 10 years
Almost 100X increase!

Source:
Thread Count Trend (Oracle T-Series CPU)

Oracle T3 (Rainbow Falls) Processor

Technology: TSMC N40GP; Die Size: 376.6mm; Transistor Count: 1 Billion; Frequency: 1.65-2.0GHz; Power: 140W; Number of Pins: 833 Sig/1284 Pwr

Rainbow Falls SerDes

<table>
<thead>
<tr>
<th></th>
<th>Coherency</th>
<th>Memory</th>
<th>PCI-E</th>
<th>XAUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link-rate (Gb/s)</td>
<td>9.6</td>
<td>6.4</td>
<td>5</td>
<td>3.125</td>
</tr>
<tr>
<td># of North-bound (RX) lanes</td>
<td>14*6</td>
<td>14*4</td>
<td>8*2</td>
<td>4*2</td>
</tr>
<tr>
<td># of South-bound (TX) lanes</td>
<td>14*6</td>
<td>10*4</td>
<td>8*2</td>
<td>4*2</td>
</tr>
<tr>
<td>Bandwidth (Gb/s)</td>
<td>1612.8</td>
<td>614.4</td>
<td>160</td>
<td>50</td>
</tr>
</tbody>
</table>

- Total raw pin BW in excess of 2.4Tbps
- ADC-based Digital DFE are used in Coherency and Memory Interface
- Self-calibrated offset cancellation improves yield and performance of ADC and SerDes link

T3 SERDES RX Block Diagram

Snapshots from ISSCC and Hot Chip

- There is a significant tradeoff between data rate, die area and bumps per lane, and power.

- Need to consider DFT, yield, manufacturability and reliability for production silicon

<table>
<thead>
<tr>
<th>Source</th>
<th>TI</th>
<th>LSI</th>
<th>Fujistu</th>
<th>Altera</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Rate (Gbps)</td>
<td>16</td>
<td>14.03</td>
<td>12.5</td>
<td>28</td>
</tr>
<tr>
<td>Power (mW/lane)</td>
<td>235</td>
<td>410</td>
<td>348*</td>
<td>246.96</td>
</tr>
<tr>
<td>Area (mm²/lane)</td>
<td>0.47</td>
<td>0.81</td>
<td>0.76</td>
<td>0.8</td>
</tr>
<tr>
<td>Channel Loss (dB)</td>
<td>34</td>
<td>26</td>
<td>34.9</td>
<td>26</td>
</tr>
<tr>
<td>Bumps per lane</td>
<td>N/A</td>
<td>14</td>
<td>17</td>
<td>21</td>
</tr>
<tr>
<td>Bumps per macro</td>
<td>N/A</td>
<td>79</td>
<td>98</td>
<td>92</td>
</tr>
<tr>
<td># DFE Taps</td>
<td>14</td>
<td>6+4</td>
<td>1</td>
<td>N/A</td>
</tr>
<tr>
<td>Technology</td>
<td>40nm TSMC</td>
<td>40nm CMOS</td>
<td>90nm CMOS</td>
<td>28nm</td>
</tr>
<tr>
<td>Power Efficiency (mW/Gbps)</td>
<td>14.69</td>
<td>29.23</td>
<td>27.84</td>
<td>8.82</td>
</tr>
<tr>
<td>Design Efficiency (mW*mm²/Gbps/dB)</td>
<td>0.2</td>
<td>0.91</td>
<td>0.6</td>
<td>0.27</td>
</tr>
</tbody>
</table>
System Channel Loss (include package)

Insertion Loss => -35dB
Crosstalk => -45dB

Required:
- Equalization (Power/Area)
- Cancellation (Power/Area)
- Error encoding (Latency)

Loss @ Nyquist doubles as data rate doubles
What Does 20Tbps Really Mean?

To deliver 20Tbps

<table>
<thead>
<tr>
<th>Data Rate (Gbps)</th>
<th># of Signals</th>
<th>Sampling Window (ps)</th>
<th>Max stub length in PCB (mil)</th>
<th>Max channel length (inch)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>1875</td>
<td>63</td>
<td>50</td>
<td>34</td>
</tr>
<tr>
<td>20</td>
<td>1500</td>
<td>50</td>
<td>25</td>
<td>28</td>
</tr>
<tr>
<td>25</td>
<td>1200</td>
<td>40</td>
<td>12.5</td>
<td>18</td>
</tr>
<tr>
<td>30</td>
<td>1000</td>
<td>33</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>35</td>
<td>857</td>
<td>29</td>
<td>3</td>
<td>-2</td>
</tr>
<tr>
<td>40</td>
<td>750</td>
<td>25</td>
<td>1.5</td>
<td>-4</td>
</tr>
</tbody>
</table>

Disclaimer: this table is for illustration purpose. The max channel length listed is first order theoretical estimation. In reality, that will determined by SERDES front end bandwidth, package, connector, xtalk and many other factors.

Note: This table assumes using the same, low loss, high cost material from 16 to 40Gbps and that the material does not improve between data points.
What does 20Tbps mean – IO Area

Area required for 20Tbps ~200-400mm²

<table>
<thead>
<tr>
<th>Data Rate (Gbps)</th>
<th># of Lanes</th>
<th>Area @ 0.80mm²/lane</th>
<th>Area @ 0.4mm²/lane</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>938</td>
<td>750</td>
<td>375</td>
</tr>
<tr>
<td>20</td>
<td>750</td>
<td>300</td>
<td>150</td>
</tr>
<tr>
<td>28</td>
<td>600</td>
<td>240</td>
<td>120</td>
</tr>
<tr>
<td>30</td>
<td>500</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>35</td>
<td>429</td>
<td>172</td>
<td>86</td>
</tr>
<tr>
<td>40</td>
<td>375</td>
<td>150</td>
<td>75</td>
</tr>
</tbody>
</table>

28 GSERDES ~200-240mm²

Area Required for 20Tbps off-chip Bandwidth

*Not all combinations of data rate and area are feasible

Source:

Global Standards for the Microelectronics Industry
What Does 20Tbps Mean – IO Power

Power

- 56Gbps ADC+DSP[3]: ADC power 2W per channel; >40mW/Gbps
- 28Gbps SERDES [2]: 8.82mW/Gbps
- 16Gbps SERDES [1]: 15mW/Gbps

Power Required for 20Tbps (10Tbps bi-di) > 200 watts!

<table>
<thead>
<tr>
<th>Power Efficiency (mW/Gbps)</th>
<th>Power (W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>120</td>
</tr>
<tr>
<td>8</td>
<td>240</td>
</tr>
<tr>
<td>12</td>
<td>360</td>
</tr>
<tr>
<td>16</td>
<td>480</td>
</tr>
<tr>
<td>20</td>
<td>600</td>
</tr>
</tbody>
</table>

Current electrical solution

Source:
The Demand for On-Chip Power

Current = \(10^{0.0913554 \times \text{year} - 182.172}\)

The Demand for Package Pin Count

Mean Values

Total Pin Count = $10^{0.0476779 \text{year} - 92.6952}$

Legend
- Low-end
- High-end
- Microprocessor

Source:
The Real Question

- Bandwidth demands continue to increase
- Channel lengths continue to decrease
- Power and Current continue to grow
- Package pins are limited

How do we continue building balanced high-performance systems?
The Paradigm Shift – A Revolution

- Do we still need mega bandwidth off a single package?
- Lots of power is consumed by moving data through a PCB channel
  - Equalization, Noise cancellation, Clk-Data-Recovery and etc.
- How about “Shrinking Everything Down”?
  - 3D Integration / System-In-Package
  - MCM, (Through Silicon Via), SSI (Stacked Silicon Interconnect), Wide-IO, Etc.
Potential 2D Electrical Solutions

- How might MCMs and Silicon interposers address the electrical problem?
  - They enable much shorter chip-to-chip interconnect
  - Shorter path length decreases power required to overcome loss which can be used to increase bandwidth
- How do we design an MCM or interposer that works for high performance applications?
## Potential MCM Configurations

<table>
<thead>
<tr>
<th>MCM Configuration</th>
<th>Internal package signals</th>
<th>Package signal pins</th>
<th>Package Total Pins</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 cpu</td>
<td>0</td>
<td>1308</td>
<td>2117</td>
</tr>
<tr>
<td>2 cpu</td>
<td>336</td>
<td>2280</td>
<td>3730</td>
</tr>
<tr>
<td>4 cpu</td>
<td>1008</td>
<td>4224</td>
<td>5444</td>
</tr>
<tr>
<td>1 cpu + local memory</td>
<td>384</td>
<td>924</td>
<td>1919</td>
</tr>
</tbody>
</table>

Example assumes 4-way processor
Yield for Multi-Chip Product

Package Yield

Relative Package Cost

70% CPU/90% Memory Yield Assumed
95% CPU/99% Memory Yield Assumed
2D Summary

• Two of the most expensive parts in high performance systems are processors and memory.

• Steps to enable the integration of multiple CPUs or CPU with memory:
  - Develop substrates and/or interposers that can carry 100s of amps and 1000s of high-speed signals.
  - Integrate more efficient cooling.
  - Decouple the yields of CPU(s) and memory without significant overhead in package or silicon interposer complexity.
The Memory Wall – Welcome to 3D

- DRAM density has outpaced bandwidth by ~75 times over the last 30 years
- Main Memory BW is limiting performance of future designs
- At the high end, the socketed DIMMs form factor needs to be replaced with direct attach 3D chip stacks supporting higher bandwidth memory interfaces
Jedec DRAM Memory Roadmap

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Process</strong></td>
<td>3x nm</td>
<td>2x_H</td>
<td>2x_L</td>
<td>1x_H</td>
<td>1x_M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDR3</td>
<td>1600</td>
<td>1866</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDR3L</td>
<td>1333</td>
<td>1600</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDR4</td>
<td></td>
<td></td>
<td>1866</td>
<td>2133</td>
<td>2400</td>
<td>2667</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDR4L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2400</td>
<td>2667</td>
</tr>
<tr>
<td>Device</td>
<td>2Gb</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2667</td>
<td>2932</td>
</tr>
<tr>
<td>Device</td>
<td></td>
<td>4Gb</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3200</td>
</tr>
<tr>
<td>Device</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8Gb</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Device</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16Gb</td>
</tr>
<tr>
<td>DIMM</td>
<td>8GB</td>
<td>16GB</td>
<td>32GB</td>
<td>64GB</td>
<td>64GB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>128GB</td>
</tr>
<tr>
<td>3DS/TSV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DDR4_2H, DDR4_4H, DDR4_8H</td>
</tr>
</tbody>
</table>

Note:
- DRAM speed: device raw speed, in Mbps.
- DIMM density: sweet spot density.
- 3x = 30-39 nm, 2xH = high 20’s nm, 1xH = high teen nm, 1xM= mid teen nm
- DDR4L: 1.0V, TBD.
Memory Latency & Bandwidth

Latency Lags Bandwidth

- Trying to overcome/cover up for lagging memory bandwidth cause added memory subsystem latency

Source:
How to Further Improve Performance

Improve bandwidth and latency through memory integration

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Configuration</th>
<th>Performance Improvement [1,2,3]</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM in DIMMs</td>
<td>8GB-32GB/core</td>
<td>1X</td>
</tr>
<tr>
<td>DRAM stacked with CPU</td>
<td>4GB/core, stacked 12 chips + CPU</td>
<td>1.5-1.9x</td>
</tr>
<tr>
<td>DRAM stacked with CPU as cache</td>
<td>4GB/core, stacked 12 chips + CPU</td>
<td>1.5x-2.8x</td>
</tr>
</tbody>
</table>

Source:
Potential 3D Technology

- Memory has demonstrated large stacks, but which stack height makes sense for high-performance memory?

- Stacking with CPU involves it’s own challenges
  - How do we handle 100’s of Amps and 1000’s of high-speed IO delivered and dispersed throughout the stacks
  - How do we develop assembly processes and methods to ensure that processors are attached and “known good” before memory stacks are attached
  - How do we develop assemble processes and methods to deal with a processor which sits on the top of the stack (enabling heat sink attach/cooling)
  - How do we ensure in the end a robust assembly

- How does the industry work together to solve these problems?
  - Standard interfaces like WideIO and others?
  - How is this compatible with custom designed silicon?
TSV - DRAM

Improve power efficiency and IO bandwidth with TSV 3D integration
- Remove redundant circuitry
- Reduce interconnect length

Source:
SSI with TSV - FPGA

“Stacked silicon interconnect provides multi-Terabit-per-second die-to-die bandwidth through 10,000 device-scale connections”

Huge bandwidth is provisioned to connect it to outside world...

Peak Off-Package Bandwidth: 2.244Tbps

Peak Serial Bandwidth: 2.784Tbps

Source:
2. Xilinx, “7 Series FPGAs Overview”, v1.6, Mar 2011
Memory Looking Forward

- DDR5 potential features:
  - Stacking technology: 3D stacking + TSV.
  - New memory architecture: revolutionary vs evolutionary
  - High speed interface with SOCKETLESS approach
  - Highly integrated packaging
  - Hybrid memory cube: Memory+Interface Logic
  - Continue to grow speed and density
Current JEDEC activity:

• DDR4 and DIMMs:
  • Rev 1 spec: Q1 C12.
  • RDIMM and LRDIMM: Mid 2012.
  • 3DS_TSV stack up to 8 height.

• Wide I/O with TSV:
  • For Mobile, smartphone.
  • Being pushed out.

• HBM [High Bandwidth Memory]:
  • 512 bits I/O
  • TSV stacking for Graphic application.
  • At early stage, behind 3DS_TSV.
JEDEC HBM:

1. Off Die Wire length are Minimum
2. Connection are independent with Memory future die size changes

- Phase 1: focus on device architecture with wide I/O
- Phase 2: 3DS+TSV stacking
- Phase 3: 3DS cube + Logic.

Global Standards for the Microelectronics Industry
Future 3DS memory Possibilities:

- Memory cube on top of LOGIC/CPU/GPU
- Memory cube next to Logic, p-t-p connection
- Traditional socket approach will not work above 4000Mbps
MCM and 3D Stacked Memory

- Flexibility of either chip scale package or KGD(s)
- Leverage a standard interface to maximize design flexibility
MCM & 3D Technology

Future Memory Stacks

Source:
[1] Sandia UHCP Consortium Academic partners include Louisiana State University, University of Illinois at Urbana-Champaign, University of Notre Dame, University of Southern California, University of Maryland, Georgia Institute of Technology, Stanford University and North Carolina State University.
State of the Union

• DDR4 is ready to kick off:
  • ES in 2012, RDIMM will be ready as well
  • Final spec and Device model will be ready by Q1C12

• Application challenges:
  • Speed bottleneck for 4R, 2DPC
  • 1DPC, Quad Rank: require DDP LRDIMM
  • For 2DPC system, RD/LRD mix remain as an issue

• 3DS_TSV and Memory Cube:
  • Require 3DS_TSV at speed above 2667
  • Need new technology for speed beyond 3200
  • Memory cube+Logic: the Holy Grail??

• Channel simulation is needed:
  • Identifying bottle neck in SI, timing, bus turn around
  • Exploring opportunity for performance enhancement
System-in-Package (SIP) – Off Package Bandwidth

- SIP (with TSV/SSI/MCM) greatly improves the power efficiency for device-scale interconnect within a single compute node
  - This helps semiconductor manufacture to keep up with Moore's law at device level for PC, Tablets, Mobile, embedded processors and many other applications
- However, for large-scale computer system, a massive system-level interconnect network is still required
- With Moore’s law and a 1-byte per flop assumption, if we keep scaling BW requirements, 300Tbps will be needed within the next decade...

Conclusion

• Electrical interconnect scales with Moore’s law so far
  • The state-of-art off chip bandwidth is at 2.5-to-5 Tbps
  • In the next 10 years, the off-package bandwidth will need to increase another 100X
  • The days of “free” bandwidth are over
• A cooperative effort is needed to realize an economically viable solution
• We must standardize various building blocks so we can all use them
  • Higher levels of integration
  • Higher complexity
• Jedec is one of critical driving forces towards this end!
The preceding is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle’s products remains at the sole discretion of Oracle.